

# TAPESTRY: A Blockchain based Service for Trusted Interaction Online

Yifan Yang, Daniel Cooper, John Collomosse, Constantin C. Drăgan, Mark Manulis,  
Jamie Steane, Arthi Manohar, Jo Briggs, Helen Jones, Wendy Moncur

**Abstract**—We present a novel blockchain based service for proving the provenance of online digital identity, exposed as an assistive tool to help non-expert users make better decisions about whom to trust online. Our service harnesses the digital personhood (DP); the longitudinal and multi-modal signals created through users' lifelong digital interactions, as a basis for evidencing the provenance of identity. We describe how users may exchange trust evidence derived from their DP, in a granular and privacy-preserving manner, with other users in order to demonstrate coherence and longevity in their behaviour online. This is enabled through a novel secure infrastructure combining hybrid on- and off-chain storage combined with deep learning for DP analytics and visualization. We show how our tools enable users to make more effective decisions on whether to trust unknown third parties online, and also to spot behavioural deviations in their own social media footprints indicative of account hijacking.

**Index Terms**—Distributed Ledger Technology, Identity, Online Trust, Artificial Intelligence, Interaction Design.

## 1 INTRODUCTION

ONLINE fraud and scams are sharply on the increase, costing the global economy in excess of US\$3 trillion in 2018 [65] and are often perpetrated through ephemeral false identities. Users struggle to make decisions on who to trust online, exposing themselves to risks from inappropriate over-disclosure of personal data. This motivates new techniques for determining the provenance and trustworthiness of the digital identities — people, business or services — encountered online.

This paper reports a novel blockchain-based service that harnesses the complex longitudinal and multi-modal signals within citizens' digitally mediated interactions (for example, on social media) to support safe online interactions. These signals created - photos shared, comments left, posts 'liked' etc. - weave a complex 'tapestry' reflecting our relationships, personality and identity, referred to as the 'Digital Personhood' (DP). Commodification of the DP now fuels a billion-dollar industry in which machine learning is increasingly utilised to help make sense of, and extract value from, the deluge of DP data. In this work we exploit the DP for the social good; through a platform (herein referred to as 'TAPESTRY') that empowers users to share 'trust evidence' of their DP in a granular, privacy preserving manner, in order to prove the provenance of their digital identity and so engender trust online.

We draw distinction between the problem of proving identity (authentication), and the problem tackled here, of

proving the provenance of a digital identity. Online security is typically reliant on traditional representations of identity, taking simple pseudonyms or email addresses at face value to interact with one another or digital services. We are now entering a new era in which citizens will construct a DP from childhood, comprising rich lifelong digital trails from social media and interactions with technology [40]. Those accumulated signals offer an increasingly viable way to prove the veracity of a digital identity. Leveraging the DP for this purpose poses significant challenges around signal processing, privacy, information security and infrastructure. Further challenges arise by designing the service for non-experts, who may have low levels of digital literacy - especially around numeracy. A fundamental tenet of the TAPESTRY platform is the preservation of the end-user as the owner of trust decisions; we do not wish to develop a 'trust traffic light' or trust scoring system. Rather we wish to summarise in an intuitive way the trust evidence disclosed from one user to another, in order to support strong decision making around trust using that evidence. TAPESTRY tackles these challenges through three novel technical contributions:

- 1) A **secure data architecture** combining off-chain storage of encrypted trust evidence derived from the DP, backed by an unpermissioned proof-of-work (PoW) blockchain to ensure the integrity and provenance of that evidence. The architecture incorporates a symmetric key sharing scheme, enabling granular disclosure of trust evidence. This provides users with agency over to whom evidence is disclosed, as well as control over the time periods and kinds of DP activity disclosed. (Secs. 3, 5).
- 2) A **machine learning** (ML) algorithm to irreversibly gist DP activity into compact descriptors that serve as the basic unit of trust evidence for sharing in the platform. We propose a deep neural network (DNN) to extract this evidence through a combi-

- Y. Yang, D. Cooper and J. Collomosse are with the Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, UK.
- C. C. Drăgan and M. Manulis are with the Surrey Centre for Cyber-Security (SCCS), University of Surrey, UK.
- J. Steane and J. Briggs are with the School of Art and Design, Northumbria University, UK. A. Manohar was with Northumbria University at the time this work was undertaken and is now at Brunel University, UK.
- W. Moncur is with the Duncan Jordanstone Centre for Art and Design (DJCAD), University of Dundee, UK. H. Jones was with DJCAD at the time this work was undertaken and is now at UCLan, UK.

Corresponding Author: yifan.yang@surrey.ac.uk.

Manuscript received May 9, 2019.

nation of semantic embedding and temporal modelling, enabling behavioural deviation to be detected over time. This in turn enables quantification of the regularity and temporal coherence of trust evidence which, combined with assurances over provenance and integrity from the blockchain, serves as the basis for users to make better trust decisions (Sec. 4).

- 3) A **data visualisation** technique for representing the regularity and coherence of the trust evidence disclosed by a user within a single static image. The design of the visualisation is evaluated and shown to enable non-expert users to quickly make accurate determinations of the trustworthiness of a digital identity previously unknown to them (Sec. 6).

In order to evaluate our technical prototype of the TAPESTRY service, we explore two user-centric case studies where valid trust judgements and the avoidance of either fraud or victimisation are desirable for users.

First, we explore the efficacy of TAPESTRY to help users to detect fraudulent profiles in the context of crowdfunding within the video games industry. Crowdfunding is a common vehicle by which small studios obtain financial support for new projects, and an online interaction in which investors consider the trustworthiness of pitchers as a primary factor in investment. We developed a controlled, workshop based evaluation of TAPESTRY in which the platform was used as an aid to investment decision making within a mocked-up crowdfunding scenario. In this scenario, we used TAPESTRY to visualize trust evidence derived from real-world social media footprints of games developers, and artificial profiles fabricated and curated in months prior to the study. We show that whilst TAPESTRY users do not make materially different trust decisions in terms of their accuracy (distinguishing the provenance of fake versus real identities), they are able to do so significantly more quickly using TAPESTRY leading to advantages when making decisions online in time pressured information overloaded situations.

Second, we explore the efficiency of TAPESTRY to help users detect unusual patterns of behaviour within their own social media profile pointing to unauthorized use (or account 'hijacking'). Again our goal is not to automatically raise an alarm or classify this behaviour, but to visually gist the trust evidence derived from a social media profile. When accustomed to the visual 'look and feel' of TAPESTRY visualisations of this trust evidence, we show that users can perceive deviations from the norm and so spot unusual patterns in online activities posted under their profile.

## 2 RELATED WORK

Open authentication models (e.g. OAuth2) exist for establishing cross-site login without credential sharing, relying upon a trusted identity provider (e.g. Google, Facebook) to approve access to a digital identity. In addition, there is a range of services which help to establish trust, for both named and anonymous/pseudonymous users. For example, Escrow is a contractual arrangement used within the Dark Web, facilitated via a third party, which engenders trust between buyer and seller for crypto-currency transactions [51]. However, TAPESTRY is not proposing yet

another access control solution or service for hosting digital identities or the DP, or for facilitating trusted transactions. Rather, TAPESTRY proposes an entirely new kind of service through which one may verify the trustworthiness of a digital identity through evidence derived from signals within an identity's DP.

### 2.1 Signals for Online Trust

The nature of trust is complex. It is '...a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviours of another' [16]. Importantly, this definition positions risk as naturally co-existing with trust: an individual accepts that the other party in an interaction may or may not act in the expected manner, but believes that their intentions are good. Offline, judgments about trust are informed by routinely available emotional and behavioural cues [10], [31], [59]. Online, these cues are usually absent. However, recent work indicates that there are alternate factors that may inform trust judgments [37].

In the context of our crowdfunding case study of TAPESTRY (c.f. Sec.7.2), prior work indicates that these factors include (i) herding, where (e.g.) potential investors are reassured by the behaviour of previous investors, on the assumption that if others are doing something, it must be the rational thing to do [11], [43] and (ii) social proof, where (e.g.) less-expert investors are encouraged to invest later in a campaign by the involvement of early investors who are experts in product development or financial investment [39]. A further factor is social engagement. For example, trust is generated when creators of a crowdfunding campaign provide investors with updates on positive progress towards published goals. This reassures investors and indicative of trust - increases their investment [35], [43]. Trust is also generated when creators link their social media accounts [62]: investors likely feel that the creator has nothing to hide. Although such observations exist in the literature, TAPESTRY is unique in aggregating evidence from such sources to aid the user in their decision making on trust.

### 2.2 Social Media Analytics for Trust

Use of online services now pervades society and automated profiling of the digital footprint is increasingly used for identity verification online [66]. Digital Personhood-related research increasingly involves exploring the behaviours and activities of an online user through the large volumes of data a user generates through e.g. social media posts and search histories that collectively comprise their digital footprint [41]. Such footprints play a crucial role across many digital economy services including user profiling [24], personality [44] and crowdfunding [52]. Therefore, it's vitally important to protect DP by early detecting any malicious activities in social media feeds to prevent economic or reputational loss. There exists various research methods in social media analytics for trust. Chalapathy et al. [9] and Yu et al. [69] explore abnormal behaviours detection from regular group patterns, while Kang et al. [38] use relationship graph to model social network activities to detect anomalous events. In [1], the authors use social graph and

text information to detect fraudulent comments in online review systems. Phua et al. [55] focus on structural metadata within the social activities instead of content. However, modeling users' behavioural norms in social media over longitudinal time periods, as well as visually representing this analysis to end-users, remains an open challenge that our research aims to address.

### 2.3 Trust and Identity over Blockchain

Blockchain's innovation is in its facilitation of direct transfer of unique digital property (e.g. currency, data, certificates) - previously reliant on third party intermediaries [2], [22]. This promotes trustless trust [2] whereby exchanges are 'unidirectionally' trustworthy and interpersonal trust, or trust in another online agent is replaced through the technology's functionality of transparency, codification and immutability as a cryptographic audit trail [22]. However, while Blockchain brings significant trust-related functionality for end users many questions remain, both about how to demonstrate and prove such trust to an end user, and if - and if so how - a blockchain service enables this over existing intermediaries. Elsdén et al. [22] catalogued over 200 blockchain applications and found that amongst identity management systems, most digital identities were provided by a third party (e.g. Facebook or email account) or required supplementary state-backed documentation (passport, social security numbers etc.) to prove an identity. Amongst self-sovereign digital identity Blockchain services, where an individual issues and controls their own identity, many involved biometrics (e.g. fingerprint or iris scan) supported by other personal information (email address, bank details etc.). Dunphy and Petitcolas's in-depth review of identity management models using DLT [21] also found a prevalence of reliance on intermediaries, with the authors additionally summarising current UK and EU regulatory challenges i.e. 'know your customer'; anti-money laundering; and data protection. There are additional challenges of not only supporting the demonstration of the technology's unique trust-supporting benefits but of communicating DP data in a visual form that supports intelligibility amongst non-expert users. TAPESTRY addresses this challenge through designing visualisations of trust evidence collected from social media, that communicate the coherence (and by extension, provenance) of the digital personhood without disclosing specifics about a subject's past activities. Related to TAPESTRY's ability to evidence social media activity in privacy preserving manner are distributed privacy-preserving social networks, such as Safebook [13] and early attempts to realize the functionalities of a social network using cryptographic techniques [4], [30]. The focus of these approaches have been on the privacy of the users and allowing them full control over their data. Thus, they allow the user to alter their data after previously committing to it. By contrast, TAPESTRY provides an immutable record of past social media activity that a user may share to evidence the provenance of their digital identity. The use of Blockchain to provide such a service - analogous to a de-centralised credit reference for identity - is unique to TAPESTRY and addresses emerging concerns among the general public around the risks to privacy and security of siloing data within centralised services or organisations.

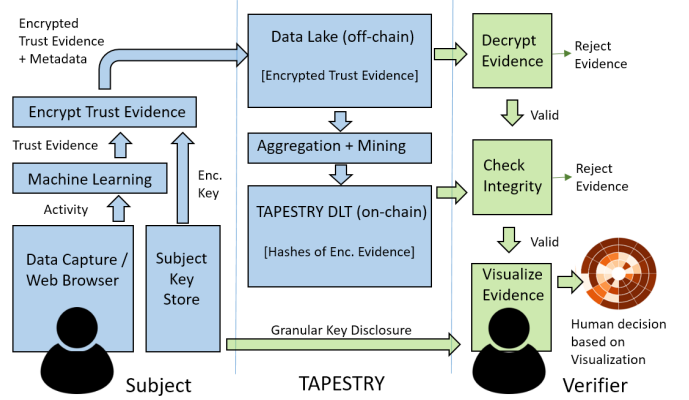


Fig. 1. TAPESTRY System architecture. Data is collected on subjects via the blue path. Digital activities (such as social media interactions) are captured on opt-in basis via a web browser extension. Trust evidence is derived from those activities via a deep neural network (DNN) and encrypted using a secret key. Keys differ between activities and change over time. Encrypted evidence is stored within a data lake, alongside metadata identifying the owner, timestamp and type of activity. A hash of that metadata serves as a unique ID to that evidence. The encrypted evidence is hashed alongside its unique ID within a proof-of-work blockchain. The green path enables a verifier to check the provenance of a subject by requesting disclosure of the relevant decryption keys. Encrypted evidence is requested from the lake; the provenance of that evidence is checked via the blockchain and it is decrypted. Analysis of the DNN signals yields a visualisation gisting the relevant period of activity that helps the user make a trust decision on the subject.

## 3 OVERVIEW OF THE TAPESTRY SERVICE

The TAPESTRY service collects, on an opt-in basis, signals derived from the digital activities of a user (the 'subject') and enables that subject to securely share those signals with another user (the 'verifier') in order to demonstrate their provenance of that subject's identity. We assume that these signals or 'trust evidence' (TE) are collected over longitudinal time periods from a rich tapestry of activities such as social media posts on various social platforms.

In a typical interaction, the verifier will determine what kinds of TE are sufficient to make a human decision on the trustworthiness of a user according to their use context. For example, on an online dating forum, or a ride-sharing service, a verifier might request evidence of a year of posting activity on any of several social media platforms in order to make their trust decision. It is a matter for the subject to decide whether to disclose the requested evidence, and indeed the act of declining creates in itself a basis for the verifying party to make a trust decision.

### 3.1 Privacy Attributes

In order to provide privacy to TAPESTRY users, TE is derived through a one-way hashing function that creates a compact, privacy-preserving gist of the semantic content of an activity (for example the text or image posted). TAPESTRY utilises a deep neural network (DNN) to perform this distillation in order to prevent content from being recovered from TE, yet enabling two pieces of TE to be compared to quantify the similarity of the content that generated it. The details of this process are described further within Sec. 4. Thus a subject may share evidence of an activity, such as a social media post, without providing the content of that post to the verifier. Furthermore, TE is stored

within the platform in an encrypted form using a secret key held by the subject. A different key is used for each TE generated from each type of activity (Facebook photo post, Twitter text post) and is changed periodically. When a subject agrees to disclose TE to a verifier they do so by sharing the relevant keys. Key generation and sharing, as well as the broader encryption scheme within TAPESTRY, is discussed in Sec. 5.

Since the volume of TE (e.g. spanning months or years) requested of a subject is typically large, TAPESTRY creates a visual gist ('visualisation') of the TE in order to make it intuitively comprehensible to the verifier. The design of the visualisation is discussed in Sec. 6. The core information communicated via the visualisation is the coherence of the user's digital history, derived from the timestamps and similarities of the TE shared by the subject. Having sight of this visualisation the verifier finally makes a human decision as to whether the user is trustworthy, in combination with other external factors such as social norms prevailing in their use context. At no point is an automated decision offered to the verifier as to the trustworthiness of the user, and at no point is the subject's decision to share TE made automatically. Rather, TAPESTRY acts as a privacy preserving conduit for the request and supply of TE.

### 3.2 De-centralised Trust Model

TAPESTRY is designed around a decentralised trust model, without reliance upon third-parties to vouch for the integrity and provenance of TE. This trust model is facilitated via a proof-of-work (PoW) Blockchain.

Recent changes in legislation (GDPR [23]) mean that users have the right to request that their personal data be deleted from any systems controlled by third-parties. In the case of TAPESTRY this meant that raw data from user activities could not be stored on-chain, as this could not be later removed. Storing only the encrypted vectors from the machine learning models, would also not comply due to the nature of the computations involved, which create an alternative digital representation of the user's behavior, thus personally identifiable information. Storing the personally identifiable information off-chain, within one or more independent data lakes (DLs), provides a method of data capture that facilitates recovering data for verification purposes, and can be deleted at the user's request.

TAPESTRY therefore uses a hybrid system of on- and off-chain storage for TE; see Fig. 1. A cloud service (of which many independently operated services are assumed to exist) maintains the DL into which the subject commits encrypted TE alongside plaintext metadata that identifies the user uniquely, as well as the timestamp and type of activity. A SHA-256 hash of the encrypted TE is stored within a proof-of-work (PoW) Blockchain, keyed by a hash of the metadata (computed also via SHA-256) which serves as a unique identifier to the TE record in the DL. In practice, a block committed to the PoW chain contains many such pairs. Fig. 2 summarises the interactions between the subject, the DL and the Blockchain during collection of TE.

On-chain hashing enables the verifier to check the provenance of TE, prior to decrypting and visualizing that evidence for human judgement. The hash of the encrypted

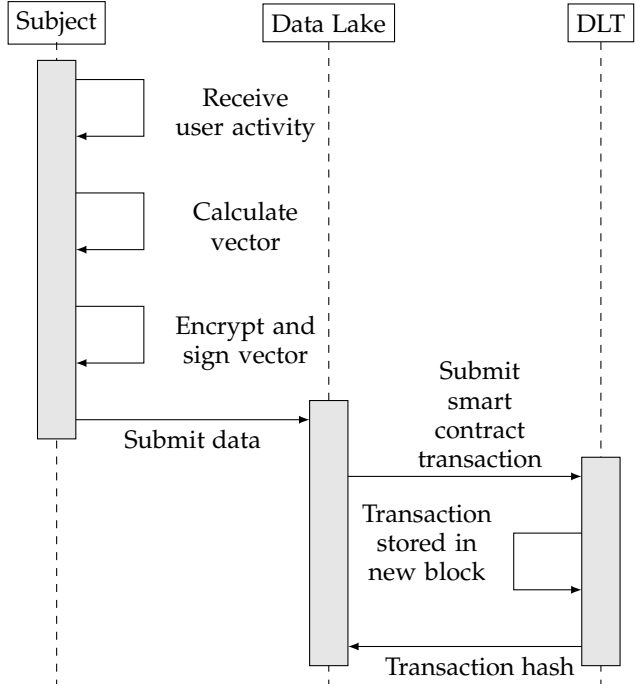


Fig. 2. Sequence diagram of the collection of trust evidence (TE) from a subject. TE is generated locally, in the form of a vector distilled from raw content via a deep neural network (DNN). The vector is encrypted, then sent to the data lake (DL) which stores the encrypted TE off-chain and records a hash of it within a new block in the PoW Blockchain (DLT).

evidence received is compared to that stored immutably within the Blockchain. This guards against an attack via fabrication of TE by the DL provider. The PoW Blockchain is implemented via Ethereum, and a smart contract to fetch (i.e. search and retrieve) and to commit (append) data to the Blockchain is provided. Failure to verify the provenance of the data, or to decrypt to the data into a parseable form (e.g. due to an invalid secret key supplied by the subject) results in an immediate rejection of the TE and strongly implies an untrustworthy interaction. Fig. 3 summarises the interactions between the verifier, subject, DL (one pictured) and the public Blockchain during sharing and verification of TE.

Given the complex security model employed to ensure secure data transmission and storage, it is plausible that user error could cause the loss of the keys, rendering their TAPESTRY data inaccessible. Users could wish to share their keys with a trusted third party key store, which could provide their keys when required or even act as a facilitator during the verification process. This optional step is analogous to sharing the private keys of cryptocurrency wallets with a centralised brokering service.

## 4 EXTRACTING TRUST EVIDENCE

We now describe the process through which activity content is hashed into trust evidence (TE). Our approach is based on the hypothesis that people have consistent (or slowly evolving) behavior and personal interests over longitudinal time periods [63], [64]. Deviations from this organic behavior pattern indicate either a non-natural (fake) account e.g. as a

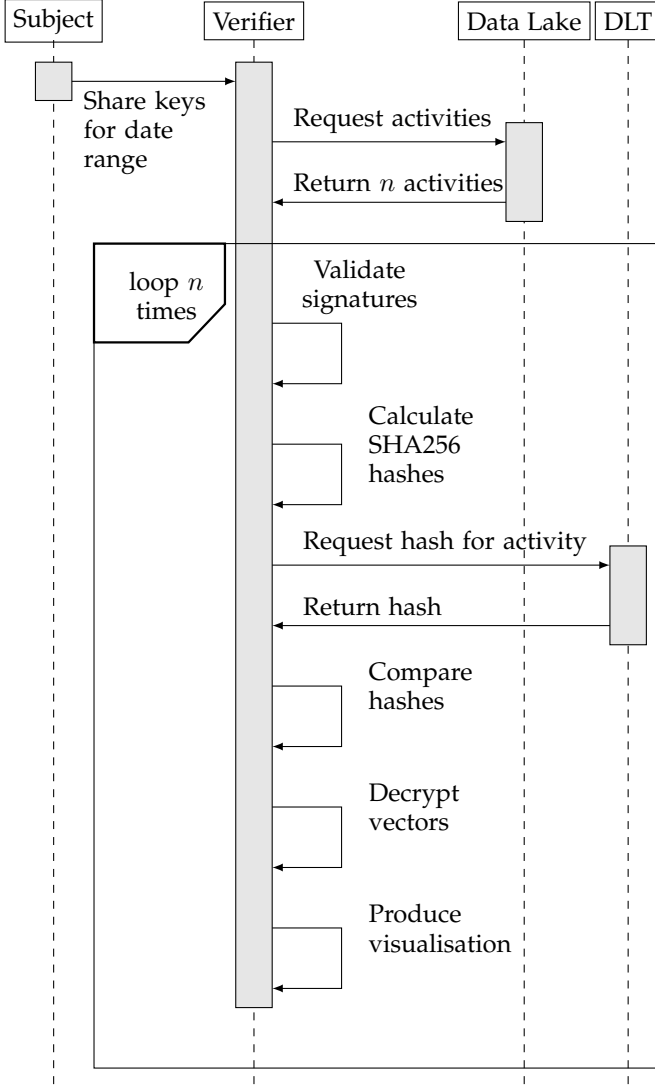


Fig. 3. Sequence diagram of the process by which a verifier determines the trustworthiness of a subject. The subject shares relevant secret keys with the verifier, enabling encrypted TE vectors to be retrieved from the data lake, verified for provenance against the Blockchain (DLT), and decrypted. The vectors are converted into a visualisation to aid the verifier's trust decision.

vehicle for spam or online scam, or a legitimate account that has been hijacked for similar purpose resulting in abnormal characteristics in the timeline.

Therefore, we tackle this deviation detection problem from behavioral norm as an outlier/anomaly detection task. Our goal is therefore to reduce the content of a post to a compact, real-valued vector (the TE) such that similar semantic content maps to similar TE. In this work, we present our deep neural network (DNN) based method to detect coherent/incoherent activities on the Twitter social media platform via analysis of text within a subject's posts. Through a combination of semantic embedding and temporal modelling, we map activity content to TE and leverage the temporal coherence of TE to help prove the provenance of a digital identity.

#### 4.1 Data pre-processing

User-generated content (UGC) on social media is often a mixture of texts, special characters, hashtags, emojis and links. This kind of raw data is not directly suitable for machine learning method. In natural language processing, a pre-processing step is necessary to clean the data to normalize it for the learning process. In this paper, we consider only meaningful texts and focus on topic analysis. We first remove special characters and retain only texts in a tweet. We then apply successive operations, including tokenisation, stop words removal and stemming, and lemmatisation.

#### 4.2 Topic Word Modeling

A word embedding is a distributed representation of words, incorporating semantic information [49] that is learned from a large corpus of text (all tweets in the collected data set in our case). Topical modeling [6] extracts a distribution of words as topics, and a distribution of topics as documents. We implemented topic word embeddings, as proposed in [47], to capture contextual information in the given document. A topic word embedding is considered as a word-topic pair  $\langle w_i, t_i \rangle$ . We considered all the tweets from one user as a document. The learned feature can enhance discrimination between words in different contexts and styles. A tweet embedding is the average of all topic word embeddings derived from the words in the tweet.

#### 4.3 Temporal Coherence via Long-Short Term Memory

The application of Deep Learning [28] is proving highly effective in making sense of signals in computer vision [42], natural language [25] and robotics [50]. We apply a DNN to learn features for each user in a temporal window, denoted as user embedding. The user embedding is regarded as a temporal pattern of tweets in a fixed time window.

The Long-Short Term Memory (LSTM) model [33] is a recurrent neural network used to model and predict time-series data. We built a sequence model to capture the coherence activities using a LSTM model and trained to extract the user's behavior norm based on their 'daily story', e.g. as played out on social media or through other online activity. We implemented a bi-directional LSTM to model the temporal coherence on a daily and weekly basis across the captured Twitter data (temporal segment). We adopted a two-layer bidirectional LSTM, followed by two fully connected layers. The input of LSTM is the topic word embedding and the output is a daily or weekly tweets embedding.

#### 4.4 Triplet network for TE Embedding

In order to compare TE from activities over time, it is necessary to learn a metric embedding in which norms may be computed to quantify deviations in the topic word embedding over time.

Triplet DNNs have been used more broadly to learn such embeddings for information retrieval e.g. for visual search [29], [58] and we similarly perform supervised learning of the TE embedding using a triplet network strategy [61]. The objective of this network structure is to map TE within the topic word embedding to a metric embedding in which

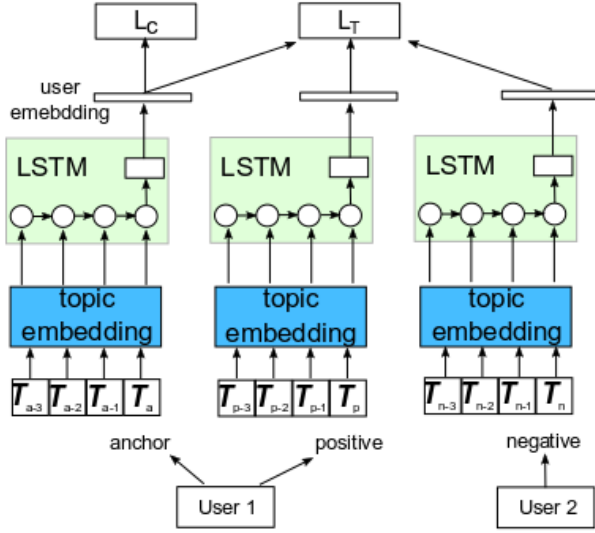


Fig. 4. Triplet LSTM architecture of our DNN used to extract TE from activity content. Content is converted to a topic embedding via averaging words within a tweet under a learned semantic (topic) model. Temporal embedding is learned via LSTM reflecting social media behavior over time. Initially the LSTM is trained as a classifier to discriminate between users within a training corpus (under cross entropy loss  $L_C$ , yielding a 'user' embedding). The embedding is then fine-tuned to yield the final TE embedding via triplet training using sequences of real twitter posts (positives), and simulated fake sequences created by splicing twitter histories together negatives), under a triplet loss  $L_T$ .

similar TE samples are pushed together and dissimilar samples pushed away from each other in the learned space. Here, similar samples are the temporal segments from one individual and dissimilar samples are the ones from different individuals. The method proved efficient in identifying different individuals from their temporal features, as learned by the prior LSTM step. We tested the method to detect compromised moments of an account, by randomly selecting a time step on one user's time-line feed. We then replaced the Tweets after the time point by the tweets from another user in order to simulate anomalous accounts, in order to provide negative exemplars for training.

Fig. 4 illustrates our network architecture and end-to-end pipeline of our machine learning algorithm for TE extraction. Given a set of  $n$  users  $\mathcal{U} = \{u_1, u_2, \dots, u_n\}$ , each user has a sequence of  $k$  tweets  $u_i = \{T_1, T_2, \dots, T_k\}$ . The tweets are first pre-processed and projected from variable length text strings to the topic embedding space. The LSTM then learns a temporal feature of the tweets for each temporal window, denoted as user embedding,  $ue_i = \{twe_1, twe_2, \dots, twe_{k-w+1}\}$ , where  $w$  is the window size. The learning strategy is a classification followed by a triplet network fine-tuning. We use a single cross-entropy loss  $\mathcal{L}_C$  for pre-training and the a combined classification and triplet loss  $\mathcal{L}_T$  based on  $L^2$  norm for fine-tuning [61]. In our experiments we use  $n = 8000$  and  $k = 800$  to train the TE embedding (c.f. Sec. 7.1).

## 5 PRIVACY AND INFORMATION SECURITY

In this section we elaborate on the data encryption and key sharing features of the TAPESTRY platform, the security properties they require and the security guarantees they

provide for the service. We follow the formalism of timeline activity proofs (TAP) [20], where we have the user's trust evidence as an activity.

### 5.1 Entities and their roles

**Subjects:** We model subjects i.e. users on whom data is collected, using secret-public key pairs  $(sk, pk)$ , with  $sk$  used generically to contain all secret information required by the user, i.e. signing keys and seed for derivation of encryption key. The public key  $pk$  is used as public identifier for the user, referred to in Sec. 3 as the user's unique ID. This is analogous to a public Bitcoin wallet address, and we allow the user to create multiple identities, and thus, hold multiple key pairs. An assumption of our system is that a physical identity can not transfer ownership or operation of such key pairs, so grounding the key pair as a basis for identity in TAPESTRY.

**Public Ledgers with external database:** For simplicity, we consider an idealized version of the blockchain approach in Section 3, where the public ledger contains an *append-only list* and an *external database*. This allows us to introduce search functionalities that verifiers require, and offer data removal to the users that have submitted their activities.

**Verifiers and Policies:** Verifiers establish *policies* - statements over different types of activities in specific intervals, that the users must satisfy. Our construction follows [20], except we consider a particular type of policy where a human verifier makes a binary decision based on the visualizations of Sec. 6.

### 5.2 Form of Trust Evidence

A user's TE maintains a strict format: the user's public identity  $pk$  (i.e. unique identified), the time it has been registered time, the type of evidence type with any optional descriptors [tags], the machine learning encoding (i.e. real-valued vector) of the evidence data  $cdata$ , and a digital signature  $\sigma$  to authenticate that it was submitted by user  $pk$ . When the user submits this trust evidence, the data component is encrypted  $cdata$ , therefore:

$$TE = \langle pk, time, type, cdata, [tags], \sigma \rangle.$$

**Building Blocks:** Our construction relies on *pseudo-random functions* (PRF) [26] and *digital signature* (DS) [15] that are *existentially unforgeable under chosen message attacks* (EUF-CMA) [27]. Additionally, we consider a *symmetric encryption scheme* (SE) with two security requirements: *indistinguishability under chosen plaintext attacks* (IND-CPA) and *wrong key detection* (WKD) [8].

**Key Management:** The trust evidence data is encrypted with a symmetric encryption scheme, where the encryption/decryption keys play an important part of the policy verification. Our solution is to derive unique encryption keys for finite time periods and for each kind of activity; in practice this could be as granular as a key per piece of TE. We realize this by assigning a random PRF seed  $s$  to each user, when they join our system. For TE, the encryption key  $ek$  is build as:

$$ek = \text{PRF}(s, \text{PRF}(s, pk, time), type).$$



For a greater degree of granularity, we may consider counting the same type of trust evidence received at the same time duration:  $\text{PRF}(s, ek, \text{count})$ . Furthermore, this allows for a granular disclosure of encryption keys only for the trust evidence required by verifiers without compromising the security of the other trust evidence.

### 5.3 Security Properties

There are two security properties that our system satisfies: *data confidentiality* that ensures the privacy of trust evidence data after the user has submitted it to the external database associated with the ledger, and *authentication policy compliance* where verifiers are only convinced by users who actually satisfy verification policies. Formal definitions are available in [20].

*Data Confidentiality:* Intuitively, this property ensures that no information is revealed concerning the trust evidence data the user is submitting, just by analyzing entries in the ledger. This property is modeled by using a probabilistic polynomial-time (PPT) adversary that is required to distinguish between two private activity encodings by seeing an entry in the database that corresponds to one of them. The entries differ only on the data component, while the public key, the time, type and tags are the same for both entries. Following the formalism of TAP, we use cryptographic primitives that satisfy the security requirements of TAP: IND-CPA for the symmetric encryption scheme, and pseudo-randomness for PRF.

*Authenticated Policy Compliance:* This property ensures that no malicious user can impersonate an honest user or fake the existence of trust evidence in the database associated to the ledger, and convince an honest verifier that they are authorized and satisfy their policy. We model this using a PPT adversary that can submit entries to the external database of a public ledger and is successful if he can convince an honest verifier to accept the evidence, when one of the two conditions are satisfied: either the adversary impersonated an honest user, or he provided a successful proof for a policy he does not satisfy. Following the formalism of TAP, we use cryptographic primitives that satisfy the security requirements of TAP: WKD for the symmetric encryption scheme, EUF-CMA for the digital signature, and pseudo-randomness for PRF.

### 5.4 Implementation details

Our cryptographic primitives are instantiated using the implementation from the python library *pynacl*. Our PRF is instantiated with the BLAKE2b [57]. In [3], it has been shown that BLAKE2b satisfies the pseudo-randomness property required by PRFs. Our DS uses the Ed25519 [5] implementation from [57] to instantiate the digital signature. Ed25519 offers existential unforgeability under chosen message attacks. Our SE is instantiated with the Salsa20 and Poly1305 MAC [36]. In [56] it has been shown that this construction satisfies IND-CPA. Moreover, the exact construction uses the technique from [8], and therefore also satisfies WKD.

## 6 MAKING TRUST EVIDENCE COMPREHENSIBLE

TAPESTRY aims to visually communicate a gist of the completeness and coherence of a subject's TE over time, so



Fig. 5. Initial designs for the TAPESTRY visualisation prototyped with the focus group. The two designs progressed for evolution and implementation in the service, based on user feedback, were the 'slash' and the 'pie' (shown second and fourth from left on the top row). The purpose of the visualisation is to communicate the completeness of TE records over the shared time period, and the coherence of activities generating that TE i.e. to flag anomalous behaviour. In many of the initial designs, these properties were reflected by spatial coverage and use of colour respectively.

that a verifier can make an informed choice as to whether to trust that subject. We rejected the security related motifs (e.g. ticks, padlocks) that are often used in online systems to signal the efficacy of a particular security function or domain of use; e.g. proportionate red-amber-green traffic light scales as used for food packaging to indicate nutritional content [14]; bronze, silver, gold hues often incorporated into badge, certificate or star rating symbolism. Such tropes convey trustworthiness as quantifiable and unequivocal (see [53]). All visualisations are persuasive to an extent [48]; this has serious design implications as TAPESTRY does not (visually) verify an online actor's trustworthiness but aims to support individuals in making their own judgments about in whom and what to trust.

Key design challenges include integrating technically and conceptually complex models into human centred approaches [18], [19], [34], [67], [68], that account for different perspectives on human understandings, values and experience [17], [22]. TAPESTRY's target user group comprises a particular demographic that is less likely to use an online service's security features [32], [54]; and has lower levels of numeracy (see [48]), than on average. Thus, there are a multiplicity of interwoven design challenges and ethical concerns around visually or otherwise communicating TE and how this is then interpreted and understood, challenges that are amplified by the non-semantic nature of TE.

### 6.1 Prototyping of Visualisations

User focus groups and lab-based workshops with user experience (UX) designers informed early concept designs for the visualisation. Our design inspirations were broad, from Knightmare, a 1990s British TV quest gameshow for children that manifested the health status of the characters using pixelated computer graphics, to more conventional



Fig. 6. Developed visualisations deployed in TAPESTRY. Left: Pie communicates interpersonal trust e.g. to help users determine whether to trust an unknown business or service online. Concentric rings of the pie correspond to different granularities of time, and shading is used to communicate coherence and volume of activity within the time period corresponding to each segment. Right: Slash communicates introspective trust e.g. to help users determine if their social media account has been compromised. Each slash corresponds to a period of time, with backslashes indicating anomalous (outlier) TE during that period. Shading is used to communicate volume of TE during that period, as with 'pie'. Six instances of the visualisations are displayed ranging from complete and coherent, to sparse TE over the time period requested by the verifier.

information and data visualisation practices. From this we then produced 12 initial designs we called 'snowflake'; 'slash'; 'radiance'; 'pie'; 'T-bar' (referring to TAPESTRY); 'T'; 'shield'; 'tiles'; 'pixel face'; 'Picasso'; 'eye' and 'pixel head' (see Fig. 5). We then rejected designs that could not depict sufficient granularity of either the completeness or coherence of TE over time. We also rejected designs that did not readily scale down (i.e. for viewing on a small screen) e.g. 'pixel head' inspired by *Knightmare*, the most anthropomorphic of the designs. We also rejected 'eye' as evocative of a surveillance system. We explored use of colour and tone, both to enable additional granularity of visual representation of the shared TE and with regard to colour's culturally situated function that could invite potentially unintended meanings for some users. Additionally, in terms of interpreting completeness of TE – within the research team it became apparent that the computer scientists associated lighter tone with more TE while designers interpreted white areas of a design as an absence of TE within a given time period. With these constraints in mind, we selected to use the idioms of 'slash' and 'pie' as the preliminary visualisations for further development.

## 6.2 Evolved Visualisation Designs

The final designs for the 'pie' and 'slash' TE visualisations are shown in Fig. 6. The choice of two designs for the visualisation reflect two use cases for deployment of TAPESTRY, evaluated in Sec. 7.

Pie is based on a simple dial that lends itself to representing temporality and accumulation of DP over days, weeks, months etc, across concentric circles, as though accumulated TE is moving towards the core of the pie. This design is used for the interpersonal trust case in which users are required to make trust decisions on an *a priori* unknown online business or service. We report of the efficacy of the visualisation in this context within the video games crowd-funding experiment of Sec. 7.2.

Slash meanwhile was intended to communicate introspective trust, where users can check TE derived from their own digital footprint (i.e. act as both subject and verifier) to determine whether their online accounts have been hacked

(Sec. 7.1.). This required a design that could visually detail sudden dissonance within an otherwise relatively uniform pattern of DP as generated over time. A visual design analogy would be a ladder in hosiery or a dropped stitch in knitwear; these draw the eye, despite their small scale, to solicit a feeling of unease in the user to invite further investigation.

Both these visualisations necessitate some initial explanation and guidance [48], though their intended meaning will require learning only once [12]. This is addressed via tutorial during the initial user sign-up to the TAPESTRY service.

## 7 EXPERIMENTAL EVALUATION

We evaluate the TAPESTRY service in two contexts. First, intraspective trust which we evaluate in the context of social media account hijacking; a user must determine whether their social account has been compromised due to anomalous posting behaviour. Second, interpersonal trust in which a user must determine the trustworthiness of another online. We evaluate this in the domain of rewards-based crowd-funding, where parties are typically unknown to one another initially and a trust judgment is fundamental to the decision to invest in a venture.

### 7.1 Intraspective Trust: Anomaly Detection

We first evaluate the efficacy of our proposed DNN approach (Sec. 4) for extracting trust evidence from social activities (text based Twitter posts). We evaluate its ability to discriminate between the behaviour of different users, and its ability to detect anomalies within the social media feed of individual users. The experiments were conducted using a public dataset of Twitter posts ('tweets') gathered by Li et al. [45] initially comprising 50 million tweets for 140 thousand users. In our experiments, we study social media footprint over longitudinal time periods and clean the data by removing users with fewer than 800 tweets in their timeline feeds. The remaining 8000 users form the basis for our experiments.



**Spot unusual behaviour using the graphics below.**

The four images below summarize a person's Twitter posting behaviour for four different time periods.

Each slash corresponds to a day. Slashes show when posts have been made. The direction of the slash shows whether the posts made that day are similar to tweets they made in the past.

1. The forward slash represents a day of tweets on topics the user normally posts about.
2. The backward slash represents a day of tweets on unusual/irregular topics for that user.
3. The absence of a slash means no tweets that day.

For each of the four images, **CLICK** the checkbox if you think the user **ACTED UNUSUALLY** in that period.

**Choose AT MOST ONE image per task** - The submission with more than one of the four images will be rejected. It is ok not to click any checkboxes if you believe there is no unusual activity.

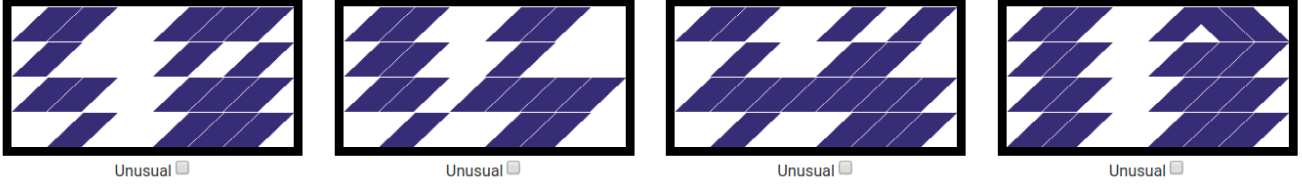


Fig. 7. Mechanical Turk experiment evaluating the efficacy of the ‘slash’ visualisation for anomaly detection (intraspective trust). MTurk workers were presented with a series of four visualisations each summarizing four weeks of trust evidence. Based on the series, users were asked which (if any) of the visualisations appeared anomalous relative to the others and so implied unusual activity was present in the social media feed.

### Evaluating Trust Evidence (TE) Embedding

We justify our choice of a LSTM to learn a temporal model for TE, via comparative evaluation against two state of the art DNN architectures; RNN and GRU. We evaluate all three architectures as a user classification problem; the networks are trained using 80% of the tweets of  $n = [500, 8000]$  users and tested on the remainder. Accuracy is measured as the number of times the system correctly identifies the user among the  $n$  possibilities. Fig. 8 shows the result of classification accuracy on all the three models as  $n$  increase. The LSTM model outperforms the other two models in most of cases.

### Evaluating Anomaly Detection

We compare the efficacy of our learned embedding at detecting anomalies within a single user’s history of TE. For this experiment we train the model on 80% of users, and testing on the remaining 20%. We compare several approaches to detecting anomalies within the test partition:

- 1) **One-class SVM (OCSVM)** [60] computes a non-linear boundary in a higher dimension space using kernel method for data project. This method allows for only positive data as ‘one class’.
- 2) **Isolation Forest (IF)** [46] evaluates the isolation degree of each data point using a random forest. This algorithm focuses on separating the outliers from the data points.
- 3) **Local outliers factor (LOF)** [7] is a distance-based method using Euclidean distance considering the density of neighborhood information.
- 4) **Proposed method.** We applied a distance based method outlier detection on our proposed user embedding. We evaluate Euclidean distance between the point and the average user embedding of first fraction of user’s activity, represented as user be-

$f_h$	Precision			Recall			F1-score		
	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
OCSVM	0.43	0.46	0.48	0.44	0.46	0.48	0.43	0.46	0.47
IF	0.71	0.62	0.57	0.75	0.64	0.59	0.73	0.63	0.58
LOF	0.12	0.24	0.34	0.13	0.25	0.37	0.12	0.24	0.35
Ours	<b>0.93</b>	<b>0.94</b>	<b>0.95</b>	<b>0.98</b>	<b>0.91</b>	<b>0.92</b>	<b>0.95</b>	<b>0.91</b>	<b>0.94</b>

TABLE 1

Evaluating the ability of our embedding to perform anomaly detection. Results compare our proposed approach to anomaly detection within the learned embedding to two common baselines.

havior norm. We consider distance larger than a selected threshold as outliers.

We simulate hijacking actions on users’ Twitter stream to evaluate anomaly detection. Given a fraction ( $f_h$ ) hijacking length, we randomly replace fraction of tweets in each user’s feed by another user’s content. These replacements are unusual behaviors that deviate from user’s activity norm. In Table 1, we show the result of precision and recall. Our proposed method outperforms all the other methods and invariant to the number of fraction number of hijacking length.

### Comprehensibility of the visualisation

In order to evaluate if our proposed visualisation is understandable and helpful for non-expert users to detect potential anomalies, we crowd-source evaluation on the Amazon Mechanical Turk (MTurk) platform. Fig. 7 depicts a representative questionnaire. We provide four successive visualisations of of a user’s activities sampled at regular intervals across their TE history. Each visualisation contains four weeks of activities. Normal behaviors are represented by forward slashes while the backward slashes as unusual behaviors detected by our proposed method in Section 4. A blank position means no activities in that time-stamps. The MTurk worker is asked to select one if they think there exists unusual behaviors among the set of four visualisations i.e.

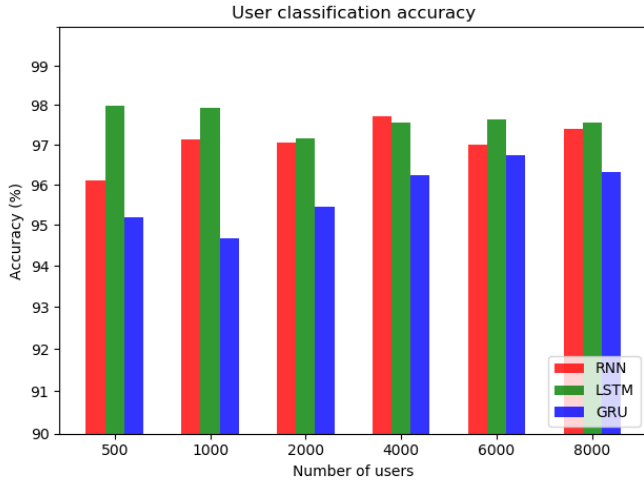


Fig. 8. Evaluating user discrimination without our DNN learned embedding for trust evidence. Our temporal modelling approach (based on LSTM) is compared to RNN and GRU sequence models. The experiment is run for  $n = [500, 8000]$  corpus of users and accuracy measured as the % of users correctly identified based on 20% of their TE.

deviation from the norm. In the experiment, we collect 1000 users in our dataset and keep users who have over 4 weeks activities in their timeline, results 537 tasks. Each task is assigned to 5 different workers. A task is correctly detected if most of workers make the correct decision. As a result, 16 workers are attributed to the whole tasks and 93.8% tasks were correctly selected by the workers. On average, each worker spend around 1 minute to finish a task including reading the instructions and submitting their decisions.

## 7.2 Interpersonal Trust: Crowdfunding

We evaluate the performance of TAPESTRY in terms of benefit to the human decision making process when establishing the trustworthiness of an individual online.

### Mock Crowd-funding Campaign

We designed an interpersonal user study comprising a decision-making task experiment presented as a mock crowdfunding campaign in the context of the video games start-up; a sector heavily reliant upon such funding. In this context, a user (the 'verifier') is invited to make an investment by another user (the 'subject') without prior knowledge of that subject. Our game was offered for investment by eight crowdfunders on a mock platform; four real games industry professionals and four fake identities that had been created two months prior to the study. We commissioned an experienced video game narrative writer who mocked-up a pitch for new video game and crowdfunding campaign. We had gained the consent of the games developers to use their real profiles in the campaign. Meanwhile the writer produced four fake profiles based on their knowledge of the gaming industry. We created fake Twitter accounts for the fake profiles and continually tweeted relevant game and entrepreneur-related comments for two months prior to the study. All eight crowdfunders had one campaign web page hosted on a password protected micro-site, which included a description of the game (constant across all candidates), a

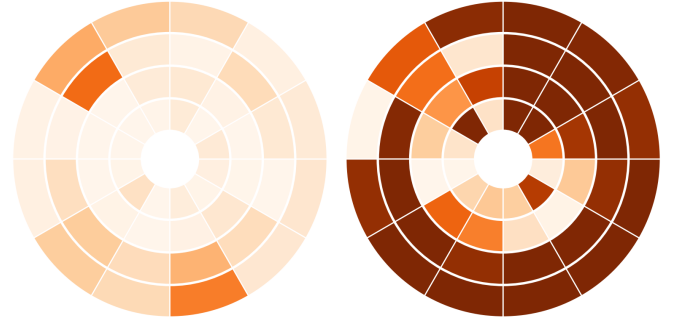


Fig. 9. Visualisation of two crowdfunders' digital footprints: strong provenance (left) and limited provenance (right). The TE requested by the 'verifier' (participant) was fixed to 2 months with concentric circles represent daily tweet activity (inner) to weeks of activity (outer).

short biography for each profile and a link to their Twitter account. The creation date of the real and fake Twitter accounts was obfuscated.

### Experimental setting

The study was run in a closed workshop with 10 participants recruited from the University campus population. After a briefing on the TAPESTRY service, participants were invited to read the crowdfunding campaigns, browse the background description and biographies and invest a hypothetical \$1000 'TAPESTRY currency' between the eight campaigns. We randomly split the participant group into two; only one group was provided with TAPESTRY visualisations on the mock crowd-funding site (Fig. 9). Participants could use the mock site or wider resources on the Internet to help them make decisions to allocate the money. The study lasted 35 minutes; participants were asked to make one decision every 5 minutes given the knowledge they gleaned from their full use of Internet resources.

### Experiment results

We evaluated the participants' performance based on their investment results, comparing the amounts invested in real and fake profiles for both groups. We consider investment in a fake profile (i. e. scammer) a bad investment. Fig. 10 shows that the accuracy of the investment results correlates to the time taken in background research; the more participants gathered information from their searches on the Internet, the more accurately they made their investment. Given the time limit, the TAPESTRY group used the visualisation tool to quickly understand the games developers' Twitter identity, speeding up their search to establish legitimacy. We can conclude that although participants reached similar, correct decisions (in terms of discriminating their investment between genuine and fake developers) the time-to-task was considerable shorter (approx. by half) for TAPESTRY users.

## 8 CONCLUSION

We presented TAPESTRY; a novel Blockchain based service that enables users to determine the provenance of online identity from their digital personhood (DP) in order to make better decisions on who to trust online. We applied the TAPESTRY service to two tasks; determining the trustworthiness of another unknown individual (interpersonal

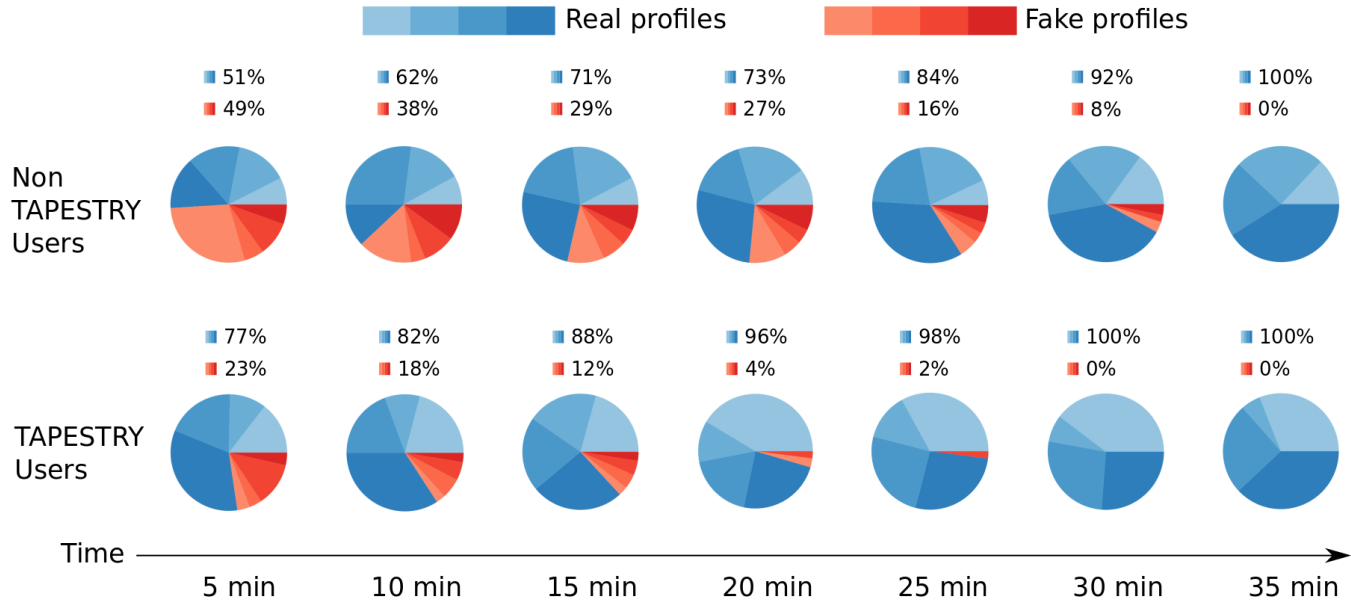


Fig. 10. Participants' investment performance in the mock crowd-funding experiment: top row is the control group without TAPESTRY visualisation; bottom row with TAPESTRY. The users with access to TAPESTRY achieved similar or better accuracy in detecting fake profiles, but did so at least twice as quickly.

trust), and determining the integrity of one's own social media feed (intraspective trust). We used machine learning techniques to extract trust evidence from social media activities in a privacy preserving manner, and a proof-of-work Blockchain to store hashes of that evidence in order to underwrite its provenance. Our service enabled users to then selectively disclose trust evidence to one another in order to prove the provenance of their identity. To improve comprehension of the high volumes of evidence shared between users, we designed visualization techniques to summarise that evidence. We evaluated the end-to-end system using a mocked up crowd-funding exercise run in a user workshop, and showed that TAPESTRY enables people to make accurate trust decision faster than the control group without access to the service. We evaluated the end-to-end system for anomaly detection and showed that TAPESTRY enabled users to detect anomalies in a social media feed with accuracy of  $\sim 94\%$ .

Currently TAPESTRY is a prototype and future work will explore at-scale deployment beyond workshop settings. At scale it will become necessary to run multiple data lake services, with users distributed across different lakes. This will add value to the PoW Blockchain which will then be maintained across multiple lakes. At this stage further characterization of the performance of the TAPESTRY Blockchain should be undertaken. Nevertheless we do not believe at-scale deployment of TAPESTRY is necessary to demonstrate the value in our hybrid on- and off-chain architecture for identity provenance, and the novel machine learning and visualization techniques developed for the service.

## ACKNOWLEDGMENTS

TAPESTRY is funded by EPSRC Grant Ref: EP/N02799X/1 under the UKRI Digital Economy Programme.

## REFERENCES

- [1] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in online reviews by network effects. In *Seventh international AAAI conference on weblogs and social media*, 2013.
- [2] M. Andreessen. Why bitcoin matters. *New York Times*, 2014.
- [3] J. Aumasson, S. Neves, Z. Wilcox-O'Hearn, and C. Winnerlein. BLAKE2: simpler, smaller, faster than MD5. In M. J. J. Jr., M. E. Locasto, P. Mohassel, and R. Safavi-Naini, editors, *Applied Cryptography and Network Security - 11th International Conference, ACNS 2013, Banff, AB, Canada, June 25-28, 2013. Proceedings*, pages 119–135, 2013. LNCS 7954.
- [4] A. Barth, D. Boneh, and B. Waters. Privacy in Encrypted Content Distribution Using Private Broadcast Encryption. In *Financial Cryptography and Data Security '06*, pages 52–64, 2006. LNCS 4107.
- [5] D. J. Bernstein, N. Duif, T. Lange, P. Schwabe, and B. Yang. High-speed high-security signatures. *J. Cryptographic Engineering*, 2(2):77–89, 2012.
- [6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [7] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.
- [8] R. Canetti, Y. T. Kalai, M. Varia, and D. Wichs. On Symmetric Encryption and Point Obfuscation. In *TCC'10*, pages 52–71, 2010. LNCS 5978.
- [9] R. Chalapathy, E. Toth, and S. Chawla. Group anomaly detection using deep generative models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 173–189. Springer, 2018.
- [10] C. Cheshire. Online trust, trustworthiness, or assurance? *Daedalus*, 140(4):4958, 2011.
- [11] R. Cialdini. The science of persuasion. *Scientific American*, 284(2):7681, 2001.
- [12] A. Cooper, R. Reimann, and D. Cronin. *About Face 3: The Essentials of Interaction Design 3rd ed.* Wiley, Indianapolis IN., 2017.
- [13] L. A. Cutillo, R. Molva, and T. Strufe. Safebook: A privacy-preserving online social network leveraging on real-life trust. *IEEE Communications Magazine*, 47(12):94–101, 2009.
- [14] P. Desai, M. E. Levine, D. J. Albers, and L. Mamykina. Pictures worth a thousand words: Reflections on visualizing personal blood glucose forecasts for individuals with type 2 diabetes. In *Proc. CHI Conf. on Human Factors in Computing Systems*, New York, 2018. ACM.
- [15] W. Diffie and M. E. Hellman. New directions in cryptography. *IEEE Trans. Information Theory*, 22(6):644–654, 1976.
- [16] R. B. a. C. D.M. Rousseau, S.B. Sitkin. Not so different after all a cross-discipline view of trust. *Academy of Management Review*, 23:393404, 1998.

- [17] P. Dourish. *Where the Action Is: The Foundations of Embodied Interaction*. MIT press Cambridge, 2001.
- [18] G. Dove, K. Halskov, J. Forlizzi, and J. Zimmerman. Ux design innovation: Challenges for working with machine learning as a design material. In *Proc. CHI Conf. on Human Factors in Computing Systems*, New York, 2017. ACM.
- [19] G. Dove and S. Jones. Using data to stimulate creative thinking in the design of new products and services. In *Proc. CHI Conf. on Designing Interactive Systems*, New York, 2014. ACM.
- [20] C. C. Dragan and M. Manulis. Bootstrapping online trust: Timeline activity proofs. In *DPM & CBT- ESORICS 2018*, pages 242–259, 2018. LNCS 11025.
- [21] P. Dunphy and F. A. P. Petitcolas. A first look at identity management schemes on the blockchain. *IEEE Security Privacy*, 16(4):20–29, July 2018.
- [22] C. Elsdén, A. Manohar, J. Briggs, M. Harding, C. Speed, and J. Vines. Making sense of blockchain applications: A typology for hci. In *Proc. CHI Conf. on Human Factors in Computing Systems*, pages 458:1–458:14, New York, 2018. ACM.
- [23] EU. General data protection regulation 679, 2016.
- [24] G. Farnadi, J. Tang, M. De Cock, and M.-F. Moens. User profiling through deep multimodal fusion. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. ACM, 2018.
- [25] Y. Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.
- [26] O. Goldreich, S. Goldwasser, and S. Micali. How to construct random functions. *J. ACM*, 33(4):792–807, 1986.
- [27] S. Goldwasser, S. Micali, and A. C. Yao. Strong signature schemes. In D. S. Johnson, R. Fagin, M. L. Fredman, D. Harel, R. M. Karp, N. A. Lynch, C. H. Papadimitriou, R. L. Rivest, W. L. Ruzzo, and J. I. Seiferas, editors, *Proceedings of the 15th Annual ACM Symposium on Theory of Computing*, 25–27 April, 1983, Boston, Massachusetts, USA, pages 431–439, 1983.
- [28] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- [29] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *Proc. ECCV*, pages 241–257, 2016.
- [30] F. Günther, M. Manulis, and T. Strufe. Cryptographic Treatment of Private User Profiles. In *Financial Cryptography Workshops’11*, pages 40–54, 2011. LNCS 7126.
- [31] J. Hancock and J. Guillory. Deception with technology. In *The handbook of the psychology of communication technology*, page 270289. Wiley-Blackwell, 2015.
- [32] C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *New Security Paradigms Workshop*, New York, 2009. ACM.
- [33] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [34] L. E. Holmquist. Intelligence on tap: artificial intelligence as a new design material. *Interactions*, 24(4):28–33, 2017.
- [35] L. Hornuf and A. Schwenbacher. Portal design and funding dynamics. *SSRN Electronic Journal*, 2015.
- [36] R. Housley. Using ChaCha20-Poly1305 Authenticated Encryption in the Cryptographic Message Syntax (CMS). *RFC 8103*, 2017.
- [37] H. Jones and W. Moncur. The role of psychology in understanding online trust. In *Psychological and Behavioral Examinations in Cyber Security*, page 109132. IGI Global, 2018.
- [38] U. Kang, L. Akoglu, and D. H. Chau. Big graph mining for the web and social media: algorithms, anomaly detection, and applications. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 677–678, 2014.
- [39] K. Kim and S. Viswanathan. The experts in the crowd: The role of reputable investors in a crowdfunding market, 2014.
- [40] A. D. K. Orzech, W. Moncur and D. Trujillo-Pisanty. Opportunities and challenges of the digital lifespan: views of service providers and citizens in the uk. *Information, Communication Society*, 21(1):1429, 2018.
- [41] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, pages 5802–5805, 2013.
- [42] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [43] V. Kuppaswamy and B. Bayus. Crowdfunding creative ideas: The dynamics of project backers in kickstarter. *SSRN Electronic Journal*, 2013.
- [44] R. Lambiotte and M. Kosinski. Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12):1934–1939, 2014.
- [45] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *KDD*, pages 1023–1031, 2012.
- [46] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.
- [47] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun. Topical word embeddings. In *AAAI*, pages 2418–2424, 2015.
- [48] M. Correll. Ethical dimensions of visualization research. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2019.
- [49] I. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [50] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.
- [51] B. M. Tzanetakis, G. Kamphausen and R. Laufenberg. The transparency paradox. building trust, resolving disputes and optimising logistics on conventional and online drugs markets. *International Journal of Drug Policy*, 35:5868, 2016.
- [52] S. Nevin, R. Gleasure, P. O’Reilly, J. Feller, S. Li, and J. Cristoforo. Social identity and social media activities in equity crowdfunding. In *Proceedings of the 13th International Symposium on Open Collaboration*. ACM, 2017.
- [53] J. Nurse, I. Agraftiotis, M. Goldsmith, S. Creese, and K. Lamberts. Two sides of the coin: measuring and communicating the trustworthiness of online information. *J. of Trust Management*, 1(5), 2014.
- [54] Ofcom. Adults media use and attitudes. Technical report, UK Government, 2018.
- [55] C. Phua, V. Lee, K. Smith, and R. Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.
- [56] G. Procter. A Security Analysis of the Composition of ChaCha20 and Poly1305. *IACR Cryptology ePrint Archive*, 2014:613, 2014.
- [57] PyNaCl. <https://pypi.org/project/PyNaCl/>.
- [58] F. Radenović, G. Tóliás, and O. Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *Proc. ECCV*, pages 3–20, 2016.
- [59] E. Rocco. Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, volume 23, page 393404. ACM, 1998.
- [60] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.
- [61] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [62] S. Vismara. Equity retention and social network theory in equity crowdfunding. *Small Business Economics*, 46(4):579590, 2016.
- [63] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In *USENIX Security Symposium*, pages 223–238, 2014.
- [64] C. Wang and B. Yang. Composite behavioral modeling for identity theft detection in online social networks. *arXiv preprint arXiv:1801.06825*, 2018.
- [65] C. C. Whitehill. The financial cost of fraud 2018. Technical report, University of Portsmouth, 2018.
- [66] R. V. Yampolskiy and V. Govindaraju. Behavioural biometrics: a survey and classification. *International Journal of Biometrics*, 1(1):81–113, 2008.
- [67] Q. Yang, N. Banovic, and J. Zimmerman. Mapping machine learning advances from hci research to reveal starting places for design innovation. In *Proc. CHI Conf. on Human Factors in Computing Systems*, New York, 2018. ACM.
- [68] Q. Yang, A. Scuito, J. Zimmerman, J. Forlizzi, and A. Steinfeld. Investigating how experienced ux designers effectively work with machine learning. In *Proc. CHI Conf. on Human Factors in Computing Systems*, New York, 2018. ACM.
- [69] R. Yü, X. He, and Y. Liu. Glad: group anomaly detection in social media analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(2):18, 2015.





**Yifan Yang** is a Research Fellow at the Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. His research interests include but not limited to machine learning, image understanding and knowledge representation and reasoning.



**Jamie Steane** is Associate Professor At Northumbria School of Design, Newcastle upon Tyne, UK. His design research lies at the intersection of design, business and digital technology and has involved the commercial development of early interactive products and services for the creative industries, financial and educational sectors.



**Daniel Cooper** is a Research Software Developer at the Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. He manages the University of Surrey DLT testbed infrastructure, as well as the design and development of software, integrating AI technologies with DLT.



**Arthi Manohar** is a researcher and lecturer in the Department of Design, Brunel University London. Her research interests include participatory design, co-design, user-centered design and Human Computer Interaction.



**John Collomosse** is a Professor at the Centre for Vision Speech and Signal Processing, University of Surrey. His interests are in the fusion of Artificial Intelligence and Distributed Ledger Technology (DLT), and he directs the Surrey Blockchain testbed comprising several UKRI funded projects in this area.



**Jo Briggs** is Associate Professor at Northumbria School of Design, Newcastle upon Tyne, UK. Her research concerns designing tools for safer and enjoyable online interaction and investigations into and through the creative collaborative economy. She leads on design for a number of interdisciplinary projects and publishes on design and socio-technical subjects.



**Constantin Cătălin Drăgan** is a Research Fellow at the Surrey Centre for Cyber Security (SCCS) and Department of Computer Science, University of Surrey, UK. His research interests are in areas of applied cryptography, security analysis, formal methods, and privacy. Previously, he worked as Research Fellow for CNRS & INRIA, France.



**Helen Jones** is a lecturer and researcher in Experimental Social Psychology at the University of Central Lancashire, UK. Her primary research interest lies in understanding how people behave online, in particular focusing on the decision-making processes that can leave them vulnerable to cyber security threats.



**Mark Manulis** is Deputy Director of Surrey Centre for Cyber Security (SCCS) and Senior Lecturer in the Department of Computer Science, University of Surrey, UK. His research interests are in applied cryptography, network security and privacy.



**Wendy Moncur** is Interdisciplinary Professor of Digital Living at the University of Dundee. She leads the Living Digital group, which focuses on human factors in cybersecurity, online identity, personal data and delivering personal agency to users.